

RESEARCH ARTICLE

Open Access



The complete genome sequence of the African buffalo (*Syncerus caffer*)

Brigitte Glanzmann*, Marlo Möller, Nikki le Roex, Gerard Tromp, Eileen G. Hoal and Paul D. van Helden

Abstract

Background: The African buffalo (*Syncerus caffer*) is an important role player in the savannah ecosystem. It has become a species of relevance because of its role as a wildlife maintenance host for an array of infectious and zoonotic diseases some of which include corridor disease, foot-and-mouth disease and bovine tuberculosis. To date, no complete genome sequence for *S. caffer* had been available for study and the genomes of other species such as the domestic cow (*Bos taurus*) had been used as a proxy for any genetics analysis conducted on this species. Here, the high coverage genome sequence of the African buffalo (*S. caffer*) is presented.

Results: A total of 19,765 genes were predicted and 19,296 genes could be successfully annotated to *S. caffer* while 469 genes remained unannotated. Moreover, in order to extend a detailed annotation of *S. caffer*, gene clusters were constructed using twelve additional mammalian genomes. The *S. caffer* genome contains 10,988 gene clusters, of which 62 are shared exclusively between *B. taurus* and *S. caffer*.

Conclusions: This study provides a unique genomic perspective for the *S. caffer*, allowing for the identification of novel variants that may play a role in the natural history and physiological adaptations.

Keywords: Genome assembly, *Syncerus caffer*, *Bos taurus*, Genetic diversity

Background

The African buffalo (*Syncerus caffer*) is the largest bovid species in the African savannah ecosystem. Buffalo are of great ecological importance because of their role as bulk feeders in the grazing hierarchy. Due to their size they are able to process taller and coarser grasses than most other species [1], playing an important facilitative role for the smaller grazers [2]. They inhabit zones with almost all vegetation types, provided a permanent water source is present. In addition, they are an important prey species and have high economic value in the ecotourism and hunting industries [3]. The African buffalo also hosts a vast array of nematodes, pathogens and infectious diseases and plays an important role in the maintenance and transmission of economically important livestock diseases such as foot-and-mouth disease (FMD), bovine brucellosis, corridor disease and bovine tuberculosis (BTB) [4, 5]. For numerous other diseases, buffalo may act as amplifier or

incidental hosts, as is the case with ehrlichiosis, Rift Valley fever and anthrax [6]. It is assumed that the African buffalo, unlike some domestic bovids (e.g. cattle), may exhibit partial resistance to some of these diseases, highlighting the importance of understanding the genetic mechanisms at work.

Advances in methods for characterizing the genetic variation in individuals, populations and species have revolutionized ecological research. Population genetic diversity, inbreeding, hybridization, species designations, dispersal patterns and evolutionary processes are just a few of the applications of genetic data in the conservation and management of wildlife [7, 8]. One of the greatest challenges of working with non-model species is the lack of availability of genome variation data with which to design these studies [9]. The assembly of an accurate genome for important non-model study species provides an invaluable resource for research. Complete and accurate reference genome information prevents the erroneous identification of polymorphisms, and misalignments [10]. In the absence of a complete reference genome, a related species can be used as a proxy reference to facilitate the identification of various single nucleotide variants (SNVs), but the amount

* Correspondence: blindcycle@sun.ac.za

SA MRC Centre for TB Research, DST/NRF Centre of Excellence for Biomedical TB Research, Division of Molecular Biology and Human Genetics, Faculty of Medicine and Health Sciences, Stellenbosch University, Cape Town, South Africa

of sequence that can be accurately mapped may be insignificant [11].

The identification of novel genetic variants in an important species such as the African buffalo may help to provide answers to numerous genetic and other research questions, including disease susceptibility, that have heretofore been unattainable [12]. Previously, we aligned buffalo short reads obtained from high-throughput sequencing to the *B. taurus* genome assembly to facilitate SNV discovery [13]. Here, we present the first complete *de novo* assembled full-length genome for the African buffalo with an assembled N50 contig size reaching 43 kilobase pairs (kbp) and an N50 scaffold length of 2.4 Mb, which represents the first fully sequenced and *de novo* assembled sequence of the African buffalo. We carried out a number of additional analyses, including evolutionary analyses and genetic content. Our results will significantly aid in understanding the genetics of the African buffalo and contribute to the fields of molecular ecology, population genetics and disease susceptibility, ultimately supporting conservation and management efforts.

Results

Twelve mate-pair libraries with different insert sizes were prepared using DNA from a 2-year old male buffalo and sequenced to a high (60-fold) coverage on an Illumina Hi-Seq 2000 and assembled *de novo*. Following data filtration and the removal of low quality reads, a total of 242.39 Gbp of usable sequence (equating to 89.78-fold coverage of the whole genome) and an average read length of 78.67 bp was obtained (Additional file 1: Table S1). The total length of the genome assembly equated to 2.68 Gbp and a total of 1235 scaffold contigs made up 90% of the genome assembly and 97.9% of the estimated length (Additional file 2: Table S2). It was estimated that the genome

size of *S. caffer* is 2732 Mbp (Additional file 3: Figure S1). A total of 19,765 genes were predicted and of these a total of 19,296 genes were annotated to *S. caffer* while a total of 469 genes could not be annotated (Table 1). Data analysis and annotation of the non-coding RNA identified a large number of miRNA and tRNA (31,940 and 36,163 respectively). The estimated heterozygosity ratio of the sequenced buffalo was 0.6×10^{-3} (Additional file 4: Figure S2) and we estimated 1,639,766 heterozygous SNVs.

To extend a detailed annotation, the *S. caffer* genome, gene clusters were constructed using twelve additional mammalian genomes (*Felis catus*, *Rattus norvegicus*, *Pan troglodytes*, *Canis familiaris*, *Equus caballus*, *H. sapiens*, *S. scrofa*, *O. aries*, *T. truncatus*, *B. taurus*, *Copelatus ferus* and *M. musculus*). The *S. caffer* genome contains 10,988 orthologous gene clusters and a total of 7321 are shared among four species (Table 2; Fig. 1a). A total of 62 predicted gene clusters are shared exclusively between *S. caffer* and *B. taurus* and 179 are unique to *S. caffer* (Fig. 1a). It was determined that the divergence time between *S. caffer* and *B. taurus* is 5.7–9.3 million years ago (MYA) (Additional file 5: Figure S5).

To investigate signatures of selection we obtained a total of 2236 1:1 orthologous gene sets in *S. caffer*, *B. taurus*, *C. ferus*, *O. aries*, *S. scrofa* and *T. truncatus* and this was based on the gene family results. Finally, we inferred 120 genes which contain positive selected sites in buffalo. These include ubiquitin carboxyl-terminal hydrolase 26 (*UCHL26*), Interleukin 19 (*IL19*) and Cyclin B (*CCNB*). Based on the comparison of orthologous gene families among the 12 mammalian species, the *S. caffer* genome has 538 expanded and 2251 contracted gene families when compared to *B. taurus* as a common ancestor (Fig. 1b). The expanded genes were coupled to a large variety of GO terms, including G-coupled protein and olfactory receptors.

Table 1 Assembly and annotation of the *S. caffer* genome

Feature		Size	Source
SOAP <i>de novo</i> assembly		–	Supplementary Table 1
Estimated genome size (assembly and 17mer)		2732 Mb	–
N50 contigs		43. kbp	Supplementary Table 2
N50 scaffolds		2.4 Mb	Supplementary Table 2
Average GC content		0.417	Supplementary Table 2
Coding genes	a. 19,296 annotated	430.18 Mb	–
	b. 469 unannotated		
Non-coding RNA (70,595 loci)	a. 31,940 micro RNA	3.25 Mb	Data not shown
	b. 1593 small nuclear RNA	184.60 kbp	
	c. 36,163 transport RNA	2.64 Mb	
	d. 899 ribosomal RNA	93.94 kbp	
Repetitive elements (37.21%)	Tandem repeats	972.19 Mb	Supplementary Table 3, 4, 5

Table 2 Summary of gene families of *S. caffer* and twelve other mammalian genomes

Species	Total number of orthologous genes	Number of unclustered genes	Number of gene families	Number of unique families	Average number of genes per family
<i>F. catus</i>	19,440	613	9549	14	1.97
<i>R. norvegicus</i>	22,656	432	9698	22	2.29
<i>P. troglodytes</i>	18,613	484	9580	25	1.89
<i>C. familiaris</i>	19,818	771	9508	16	2.00
<i>E. caballus</i>	20,372	229	9364	22	2.15
<i>H. sapiens</i>	22,214	371	9836	24	2.22
<i>M. musculus</i>	22,484	655	9718	45	2.25
<i>S. scrofa</i>	21,526	1860	9392	120	2.09
<i>O. aries</i>	20,786	523	9878	29	2.05
<i>T. truncatus</i>	16,476	245	9082	2	1.79
<i>B. taurus</i>	19,950	108	9552	1	2.08
<i>S. caffer</i>	19,292	1295	8888	52	2.02
<i>C. ferus</i> ^a	23,017	784	9269	16	2.40

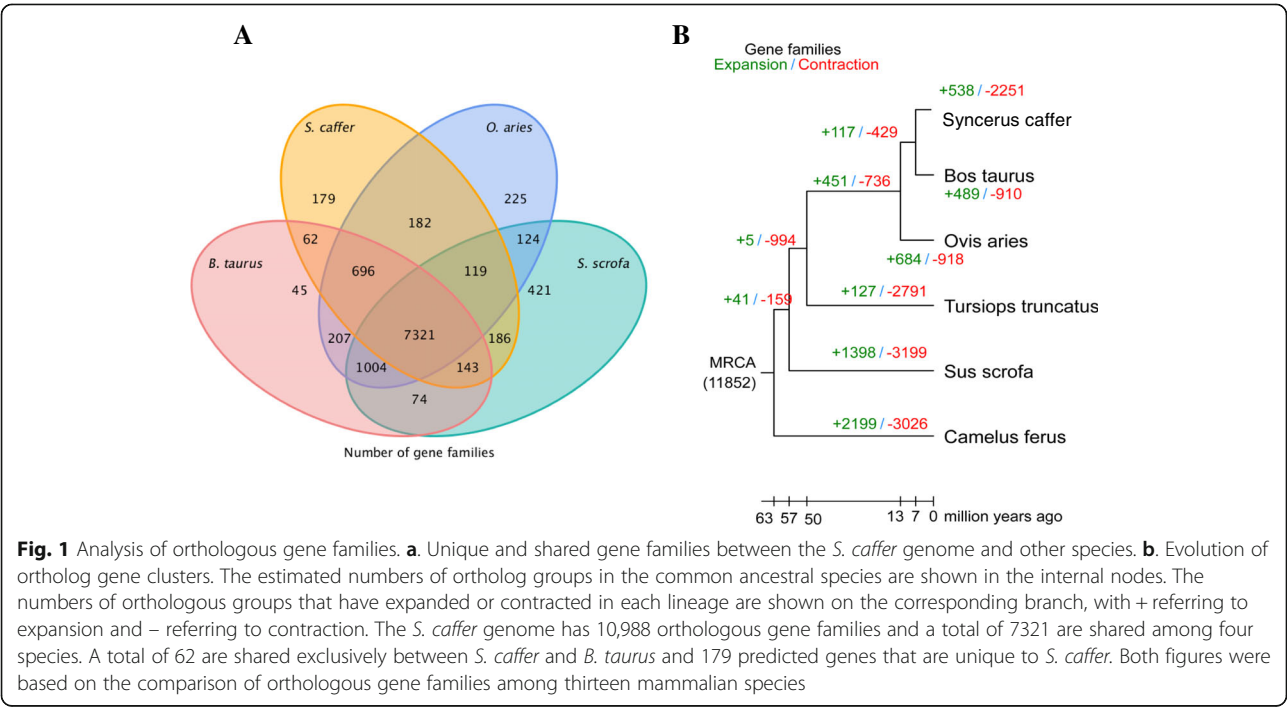
^a*C. ferus* genome was downloaded from NCBI(ftp.ncbi.nih.gov/genomes/Camelus_ferus), while all other genomes were download from Ensemble release-78(ftp.ensembl.org/pub/release-78)

The figures in bold are intended to highlight the information obtained from the buffalo

Discussion

The African buffalo has become a species of great interest in recent years because it serves as maintenance host for numerous infectious and zoonotic diseases such as FMD, corridor disease and bovine tuberculosis [5]. In addition, its high economic value in the ecotourism and trophy hunting industries make this species invaluable to game ranchers and breeders [4, 5, 14]. The African buffalo is one of only three main species of buffalo found in the world. The other two include the American

Bison (*Bison bison*) and the domesticated Asiatic buffalo (*Bubalus bubalis*). Reference genomes of 2.82 Gbp and 2.77 Gbp have been assembled for these respectively, but are not publically available [15, 16]. The *Bubalus bubalis* genome was found to encode 21,550 protein coding genes, which is comparable to the African buffalo genome we report here, which is 2.73 Gbp in size with a total of 19,765 predicted protein coding genes and of these 19,296 were annotated. Moreover, the annotation of non-coding RNA identified unexpectedly large numbers of miRNA and



tRNA [17, 18]. The use of standard tools namely tRNAscanner SE (version 1.23) as well as alignment using BLAST and INFERNAL for *de novo* non-coding RNA annotation consistently identified 31,940 miRNA genes. The large number is implausible (data not shown), suggesting the need for improved prediction tools in some species including *S. caffer*.

Although big game species such as the African lion, African elephant and the white rhino have been sequenced [19–21], the only suitable reference genome publically available for the alignment of African buffalo sequences was that of *B. taurus* [13]. *S. caffer* has 52 chromosomes [22] compared to the 30 chromosomes found in *B. taurus* [23]. Our analyses here estimated 1,639,766 heterozygous SNVs in the sequenced buffalo, compared to the 3,833,249 heterozygous variants identified in the Holstein *B. taurus* genome [24]. The African buffalo is not an antecedent of the cow and their most recent common ancestor is estimated to have existed approximately 5 - 10 MYA, at the time of divergence of the sub-tribes Bubalina, which is composed of the *Syncerus* and *Bubalus* genera and Bovina, comprising the *Bos* and *Bison* genera [25, 26]. In previous work, we determined that only 19 to 23% of the low coverage buffalo short reads mapped to the cow reference genome using BWA and Bowtie, illustrating the need for a species-specific reference [13].

We anticipate that the annotated African buffalo genome will facilitate our future genetic association studies of susceptibility to BTB, which is a threat to conservation areas in South Africa [27]. Previously, we identified novel SNVs by Sanger sequencing conserved regions across species, a process that was time-consuming and often resulted in non-specific amplification [28]. Alternatively, we had to rely on short reads mapped to an unsuitable reference genome [13]. It will now be possible to design species-specific primers for susceptibility genes of interest, based on the gene annotation completed here. Low coverage genome sequencing of additional buffalo will allow us to establish a SNV database, which would also be a resource for future population genetic and disease association studies [29].

Several of the genes that were substantially expanded and contracted in the African buffalo compared to other mammals are involved in immunity. This includes the contracted genes Chemokine (C-X-C motif) ligand 2 (*CXCL2*) and complement component 8 alpha subunit (*C8A*) as well as the expanded genes T cell receptor gamma variable 3 (*TRGV3*) and Killer immunoglobulin-like receptor KIR3DL splice variant 3 (*KIR3DL*). In total it is estimated that there are 175 genes involved in immune responses in the African buffalo and these are possible candidate genes to investigate in disease susceptibility studies.

Conclusions

In summary, the African buffalo genome offers unique insight into the phylogenetic history and adaptation of an ecologically important species. Additionally, the availability of a complete reference genome allows for improved mapping of short reads, thereby aiding in novel SNV discovery and future genetic studies.

Methods

Buffalo blood was obtained from a number of *S. caffer* individuals from the southern section of the Kruger National Park, South Africa for other projects. Based on the quantity and quality of the DNA available and to meet the sequencing requirements, the DNA from a 2-year old male was chosen for DNA sequencing. DNA was extracted from whole blood using the salt-chloroform extraction method [30]. No ethics approval was required as the DNA was collected and extracted for a previous study, under the directive of South African National Parks (SANParks), and its use in the present study is incidental.

Genome assembly for *S. caffer*

All libraries were sequenced using an Illumina Hi-Seq 2000 instrument. The sequencing libraries were constructed with insert sizes of 170 bp, 500 bp, 800 bp, 2 kbp, 5 kbp and 10 kbp respectively. *S. caffer* genome scaffolds (1235 scaffolds; N50 contig: 43 kbp, N50 scaffold: 2.4 Mb) were aligned to the reference *B. taurus* genome assembly. The protein coding genes in *S. caffer* were determined through the integration of annotations from homology-based methods as well as *de novo* gene assembly. For homology-based prediction, proteins from *B. taurus*, *Homo sapiens*, *Mus musculus*, *Ovis aries*, *Sus scrofa*, *Tursiops truncatus* were mapped to the *S. caffer* genome using TblastN [31] and were then submitted to GeneWise version 2.2.0 [32] in order to obtain gene models. For *de novo* prediction, two software programs were used: SNAP [33] and AUGUSTUS [34].

Genome size evaluation

The genome size of an individual can be estimated from the K-mer frequency of the read data. Importantly, the K should be large enough that most of the genome can be distinguished. For most eukaryotic genomes, a K-mer value of 17 is used ($K = 17$). For the present study, a total of 74 Gb (approximately 30X) of the data could be retained for 17-mer analyses. Simulations were done to estimate the heterozygosity ratio.

Genome assembly

The short reads were assembled using the latest version of SOAPdenovo (<http://soap.genomics.org.cn/soapdenovo.html>),

a genome assembler that has been developed for specific use with next generation short read sequences [35]. SOAPdenovo uses the de Bruijn graph algorithm, which is sensitive to sequencing errors and for this reason, only high-quality, filtered data were used for the *de novo* assembly. Gaps in the initial assembly were filled using Gap-closer [35]. Short reads from fragmented small insert-size libraries (<500 bp) were assembled into contigs using sequence overlap information. Contigs were not extended into regions in which repeat sequences created ambiguous associations. The resulting assembly contained a small contig N50 with a length of 43 kbp and a scaffold N50 length of 2.4 Mbp.

GC content and sequencing depth analysis

The completeness of the genome assembly was evaluated by aligning sequence reads to the newly generated genome assembly and subsequently determining the percentage of total aligned reads. High quality reads i.e. reads with a percentage of high quality bases that satisfies a user-specific cutoff, with an average coverage of 38-fold were aligned to the *S. caffer* genome using BWA-MEM [36]. A total of 99.93% reads were aligned to the assembly.

Variable GC content difference is a primary determinant for the non-random distribution of sequencing depth [37]. Moreover, the distribution of GC content vs. sequencing depth is a means of ascertaining sequencing bias or contamination (Additional file 6: Figure S3 and Additional file 7: Figure S4). Regions with very low GC contents (<20%) or very high GC contents (>80%) will have a low sequencing depth. Should specific regions of the genome have a GC content that is significantly different from that predicted, it is reasonable to assume that there may be bacterial, viral or fungal contamination in the sample and reads should be eliminated by additional alignment. For the purposes of this study, two sequencing depth distribution blocks were obtained; one with an average sequencing depth of 38X and the other with a sequencing depth of 19X. The sample that was sequenced was a male, and it is anticipated that hemizygous regions will have lower coverage depth because there is only one allelic counterpart on the Y chromosome. A total of 2825 sequences are found at low coverage (19X) (Additional file 7: Figure S4) and were subsequently aligned to the X and Y chromosomes of *B. taurus*. It was found that 99.85% of the sequences align directly to the X and Y chromosome of *B. taurus* while 0.6 and 0.64% align to the fungi and bacterial databases respectively. It was therefore concluded that this block of low coverage region forms part of the sex chromosomes and would thus not adversely affect the *de novo* assembly.

Gene annotation

1. Repetitive element annotation

Tandem Repeats Finder (TRF) was used to identify non-interspersed repetitive elements. Transposable elements (TEs) were predicted in the homology searches found in Repbase TE libraries using Repeat ProteinMask and RepeatMasker (Additional file 8: Table S3, Additional file 9: Table S4 and Additional file 10: Table S5).

2. Non-coding RNA annotation

A total of four non-coding RNA (ncRNA) types were annotated.

- micro RNA (miRNA)
- transfer RNA (tRNA)
- ribosomal RNA (rRNA)
- small nuclear RNA (sn-RNA)

All four of these RNA types were found in *S. caffer* genome using the complete genome sequence. Scanning for tRNA was performed using tRNAscan-SE [38] by using a short interspersed elements (SINEs) premasked genome to search for reliable tRNA positions. The snRNAs and miRNAs were identified by aligning with BLAST and INFERNAL to search for putative sequences in the Rfam database [39]. The rRNA fragments were identified by aligning the rRNA template sequences from the human genome using BlastN.

3. Gene prediction

The protein coding genes in *S. caffer* were determined through the integration of annotations from homology-based methods as well as *de novo* gene assembly. For homology-based prediction, proteins from *B. taurus*, *H. sapiens*, *M. musculus*, *O. aries*, *S. scrofa*, *T. truncatus* and mapped to the buffalo genomes using TblastN [31] and were then submitted to GeneWise [32] in order to obtain gene models. For *de novo* prediction, two software programs were used: SNAP [33] and Augustus [34] with gene model parameters trained from *H. sapiens*, and filtered partial genes and small genes that had less than 150 bp coding length. This followed with the alignment of the predictions to a TE protein database using BlastP [31] with an E-value $\leq 1e-5$ and filtered TE-derived genes that had more than 50% alignment rate (Additional file 11: Table S6).

Gene family construction

To determine the genetic evolution in *S. caffer*, gene cluster analysis included the genomes of 13 mammals. Proteins of all genes for each of the species chosen were analyzed using Treefam [40]. All proteins sequences were aligned to themselves using BlastP [31] with E-value cut-off of $1e-7$ (Table 2).

Phylogeny and divergence

A total of 1745 single copy gene clusters were identified and were used to construct a phylogenetic tree of the 13 mammal species. MUSCLE [3] was used for alignment and gaps removed by Gblocks [41]. A total of 2,265,138 (55.45%) remained and were used to construct the phylogeny. Moreover, the species divergence time was estimated based on 195,689 fourfold degenerate sites via Bayesian estimation approach using PAML [42]. The phylogenetic relationship of *S. caffer* and twelve additional mammals was predicted. The data can be accessed using the following link: <http://purl.org/phylo/treebase/phyloids/study/TB2:S20207?x-access-code=f3bd b11f9f55ac0609abf60ab1c01255&format=html>.

Gene family expansion and contraction

The evolutionary changes in the protein family size (expansion or contraction) were analyzed using the CAFÉ program [43]. This package assesses the protein family expansion or contraction based on the topology of the phylogenetic tree (Fig. 1).

Branch site positive selection

Orthologues were aligned using the PRANK alignment algorithm available the GUIDANCE software program [44], which can improve the performance of positive selection inference by filtering out unreliable alignment regions.

Ka/Ks of *S. caffer* and *B. taurus*

The ratio of the total number of non-synonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks), referred to as Ka/Ks, can be used as an indicator of selective pressure acting on a protein-coding gene. In total, 19,994 orthologues from both *S. caffer* and *B. taurus* were chosen using Reciprocal Best Hits (RBH) methodology [45] based on the BLAST alignment. Subsequently, the coding regions of the two species were aligned using webPRANK [46] and unreliable regions were removed. Ka/Ks scores were calculated using the KaKs Calculator [47].

Additional files

Additional file 1: Table S1. Data statistics for de novo assembly of the *S. caffer* genome. (PDF 63 kb)

Additional file 2: Table S2. Genome assembly results. (PDF 62 kb)

Additional file 3: Figure S1. 17-mer coverage distribution. (PDF 69 kb)

Additional file 4: Figure S2. Simulations to estimate heterozygosity ratio. (PDF 72 kb)

Additional file 5: Figure S5. Phylogeny and divergence of 13 mammals. (PDF 352 kb)

Additional file 6: Figure S3. Distribution of the read depth for the de novo assembled *S. caffer* genome. (PDF 86 kb)

Additional file 7: Figure S4. GC content and average sequencing depth values for the *S. caffer* genome. (PDF 361 kb)

Additional file 8: Table S3. Summary of repetitive elements obtained using 3 pipelines. (PDF 46 kb)

Additional file 9: Table S4. Summary of TE classification. (PDF 49 kb)

Additional file 10: Table S5. Top 10 copy number variations TE subfamily in African buffalo genome. (PDF 50 kb)

Additional file 11: Table S6. Summary of function annotation for African buffalo. (PDF 45 kb)

Abbreviations

BTB: Bovine tuberculosis; FMD: Foot and mouth disease; kbp: kilobase pairs; MYA: Million years ago; RBH: Reciprocal best hits; SANParks: South African National Parks; SINE: Short interspersed elements; SNVs: Single nucleotide variants; TE: Transposable elements; TRF: Tandem repeats finder.

Acknowledgements

We gratefully acknowledge the Kruger National Park and SANParks for providing the sample used in this study. The South African Medical Research Council, the South African National Research Foundation and Stellenbosch University supported this work financially.

Funding

This project was funded by the National Research Foundation of South Africa, the South African Medical Research Council and Stellenbosch University.

Availability of data and materials

The complete genome sequence has been submitted to the National Centre for Biotechnology Information with accession number PRJNA341313 and can be accessed through the website.

Authors' contributions

NLR, EH and PVH conceived the project. NLR and PVH were responsible for sample collection and DNA extraction was carried out by NLR. GT and BG were responsible for the bioinformatics analysis of the data. BG and MM wrote the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval

Authorization was provided by Stellenbosch University Animal Care and Use (SU ACU) and conducted in accordance with the policies and procedures as stipulated by South African National Parks (SANParks).

Received: 14 April 2016 Accepted: 2 December 2016

Published online: 07 December 2016

References

1. Michel AL, Bengis RG. The African buffalo: a villain for inter-species spread of infectious diseases in Southern Africa. *Onderstepoort J Vet Res.* 2012;79:453.
2. Jolles AE. Population biology of African buffalo (*Syncerus caffer*) at Hluhluwe-iMfolozi Park. *South Africa Afr J Ecol.* 2007;45:398–406.
3. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.* 2004;5:113.
4. le Roex N, Koets AP, van Helden PD, Hoal EG. Gene polymorphisms in African buffalo associated with susceptibility to bovine tuberculosis infection. *PLoS One.* 2013;8:e64494.
5. Grobler JP, Van Der Bank FH. Genetic diversity and isolation in African Buffalo (*Syncerus caffer*). *Biochem Syst Ecol.* 1996;24:757–61.
6. Beechler BR, Bengis R, Swanepoel R, Paweska JT, Kemp A, van Vuren PJ, et al. Rift valley Fever in Kruger national park: do buffalo play a role in the inter-epidemic circulation of virus? *Transbound Emerg Dis.* 2015;62:24–32.
7. Kirk H, Freeland JR. Applications and implications of neutral versus non-neutral markers in molecular ecology. *Int J Mol Sci.* 2011;12:3966–88.

8. Shafer ABA, Wolf JBW, Alves PC, Bergström L, Bruford MW, Brännström I, et al. Genomics and the challenging translation into conservation practice. *Trends Ecol Evol*. 2015;30:78–87.
9. Garvin MR, Saitoh K, Garret AJ. Application of single nucleotide polymorphisms to non-model species: a technical review. *Mol Ecol Resour*. 2010;10:915–34.
10. Baker M. De novo genome assembly: what every biologist should know. *Nat Methods*. 2012;9:333–7.
11. Everett MV, Grau ED, Seeb JE. Short reads and nonmodel species: exploring the complexities of next-generation sequence assembly and SNP discovery in the absence of a reference genome. *Mol Ecol Resour*. 2011;11:93–108.
12. le Roex N, van Helden PD, Koets AP, Hoal EG. Bovine TB in livestock and wildlife: what's in the genes? *Physiol Genomics*. 2013;45:631–7.
13. le Roex N, Noyes H, Brass A, Bradley DG, Kemp SJ, Kay S, et al. Novel SNP Discovery in African Buffalo, *Syncerus caffer*, using high-throughput Sequencing. *PLoS One*. 2012;7:e48792.
14. Booth VR. A comparison of the prices of hunting tourism in Southern and Eastern Africa. Budapest: International Council for Game and Wildlife Conservation and Food and Agriculture Organisation of the United Nations; 2009.
15. Dobson LK. Sequencing the Genome of the North American Bison [Internet]. [Thesis]. 2015 [cited 2016 Jul 13]. Available from: <http://oaktrust.library.tamu.edu/handle/1969.1/155759>
16. Water buffalo genome sequenced [Internet]. [cited 2016 Jul 14]. Available from: <http://www.cabi.org/animalscience/news/23623>
17. Iben JR, Maraja RJ. tRNAomics: tRNA gene copy number variation and codon use provide bioinformatic evidence of a new anticodon:codon wobble pair in a eukaryote. *RNA*. 2012;18:1358–72.
18. Ghorai A, Ghosh U. miRNA gene counts in chromosomes vary widely in a species and biogenesis of miRNA largely depends on transcription or post-transcriptional processing of coding genes. *Front. Genet*. [Internet]. 2014 [cited 2016 Jul 13];5. Available from: <http://journal.frontiersin.org/article/10.3389/fgene.2014.00100/abstract>
19. Elephant (*Loxodonta africana*) Genome Browser Gateway [Internet]. [cited 2016 Feb 19]. Available from: <http://genome.ucsc.edu/cgi-bin/hgGateway?org=Elephant&db=loxAfr3&hgid=154866831>
20. Pre Ensembl 78: *Ceratotherium simum simum* - Description [Internet]. [cited 2016 Feb 19]. Available from: http://pre.ensembl.org/Ceratotherium_simum_simum/info/Index
21. Cho YS, Hu L, Hou H, Lee H, Xu J, Kwon S, et al. The tiger genome and comparative analysis with lion and snow leopard genomes. *Nat Commun*. 2013;4:2433.
22. Smits N, Berthouly C, Cornélis D, Heller R, Van Hooft P, Chardonnet P, et al. Pan-African Genetic Structure in the African Buffalo (*Syncerus caffer*): Investigating Intraspecific Divergence. *PLoS ONE* [Internet]. 2013 [cited 2016 Jul 14];8. Available from: <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3578844/>
23. Qiu Q, Zhang G, Ma T, Qian W, Wang J, Ye Z, et al. The yak genome and adaptation to life at high altitude. *Nat Genet*. 2012;44:946–9.
24. Köks S, Lilleoja R, Reimann E, Salumets A, Reemann P, Jaakma Ü. Sequencing and annotated analysis of the Holstein cow genome. *Mamm Genome Off J Int Mamm Genome Soc*. 2013;24:309–21.
25. Bovine Genome Sequencing and Analysis Consortium, Elsik CG, Tellam RL, Worley KC, Gibbs RA, Muzny DM, et al. The genome sequence of taurine cattle: a window to ruminant biology and evolution. *Science*. 2009;324:522–8.
26. MacEachern S, McEwan J, Goddard M. Phylogenetic reconstruction and the identification of ancient polymorphism in the Bovini tribe (Bovidae, Bovinae). *BMC Genomics*. 2009;10:177.
27. le Roex N, Cooper D, van Helden PD, Hoal EG, Jolles AE. Disease Control in Wildlife: Evaluating a Test and Cull Programme for Bovine Tuberculosis in African Buffalo. *Transbound Emerg Dis*. 2015;63:647–57.
28. Müller A, Möller M, Adams LA, Warren RM, Hoal EG, van Helden PD. Comparative analysis of a putative tuberculosis-susceptibility gene, MC3R, and pseudogene sequences in cattle, African buffalo, hyena, rhinoceros and other African bovids and ruminants. *Cytogenet Genome Res*. 2012;136:117–22.
29. le Roex N, Berrington CM, Hoal EG, van Helden PD. Selective breeding: the future of TB management in African buffalo? *Acta Trop*. 2015;149:38–44.
30. Miller SA, Dykes DD, Polesky HF. A simple salting out procedure for extracting DNA from human nucleated cells. *Nucleic Acids Res*. 1988;16:1215.
31. Mount DW. Using the Basic Local Alignment Search Tool (BLAST). *CSH Protoc*. 2007;2007:pdb.top17.
32. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. *Genome Res*. 2004;14:988–95.
33. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*. 2003;19:215–25.
34. Korf I. Gene finding in novel genomes. *BMC Bioinformatics*. 2004;5:59.
35. Li R, Li Y, Kristiansen K, Wang J. SOAP: short oligonucleotide alignment program. *Bioinformatics*. 2008;24:713–4.
36. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinforma Oxf Engl*. 2009;25:1754–60.
37. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*. 2008;456:53–9.
38. Lowe TM, Eddy SR. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res*. 1997;25:0955–64.
39. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. *Nucleic Acids Res*. 2003;31:439–51.
40. Li H, Coglan A, Ruan J, Coin LJ, Hériché J-K, Osmotherly L, et al. TreeFam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Res*. 2006;34:D572–80.
41. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst Biol*. 2007;56:564–77.
42. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007;24:1586–91.
43. De Bie T, Cristianini N, Demuth JP, Hahn MW. CAFE: a computational tool for the study of gene family evolution. *Bioinforma Oxf Engl*. 2006;22:1269–71.
44. Penn O, Privman E, Ashkenazy H, Landan G, Graur D, Pupko T. GUIDANCE: a web server for assessing alignment confidence scores. *Nucleic Acids Res*. 2010;38:W23–8.
45. Ward N, Moreno-Hagelsieb G. Quickly Finding Orthologs as Reciprocal Best Hits with BLAT, LAST, and UBLAST: How Much Do We Miss? *PLoS ONE*. 2014;9:e101850.
46. Löytynoja A, Goldman N. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. *BMC Bioinformatics*. 2010;11:579.
47. Zhang Z, Li J, Zhao X-Q, Wang J, Wong GK-S, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics*. 2006;4:259–63.

Submit your next manuscript to BioMed Central and we will help you at every step:

- We accept pre-submission inquiries
- Our selector tool helps you to find the most relevant journal
- We provide round the clock customer support
- Convenient online submission
- Thorough peer review
- Inclusion in PubMed and all major indexing services
- Maximum visibility for your research

Submit your manuscript at
www.biomedcentral.com/submit

